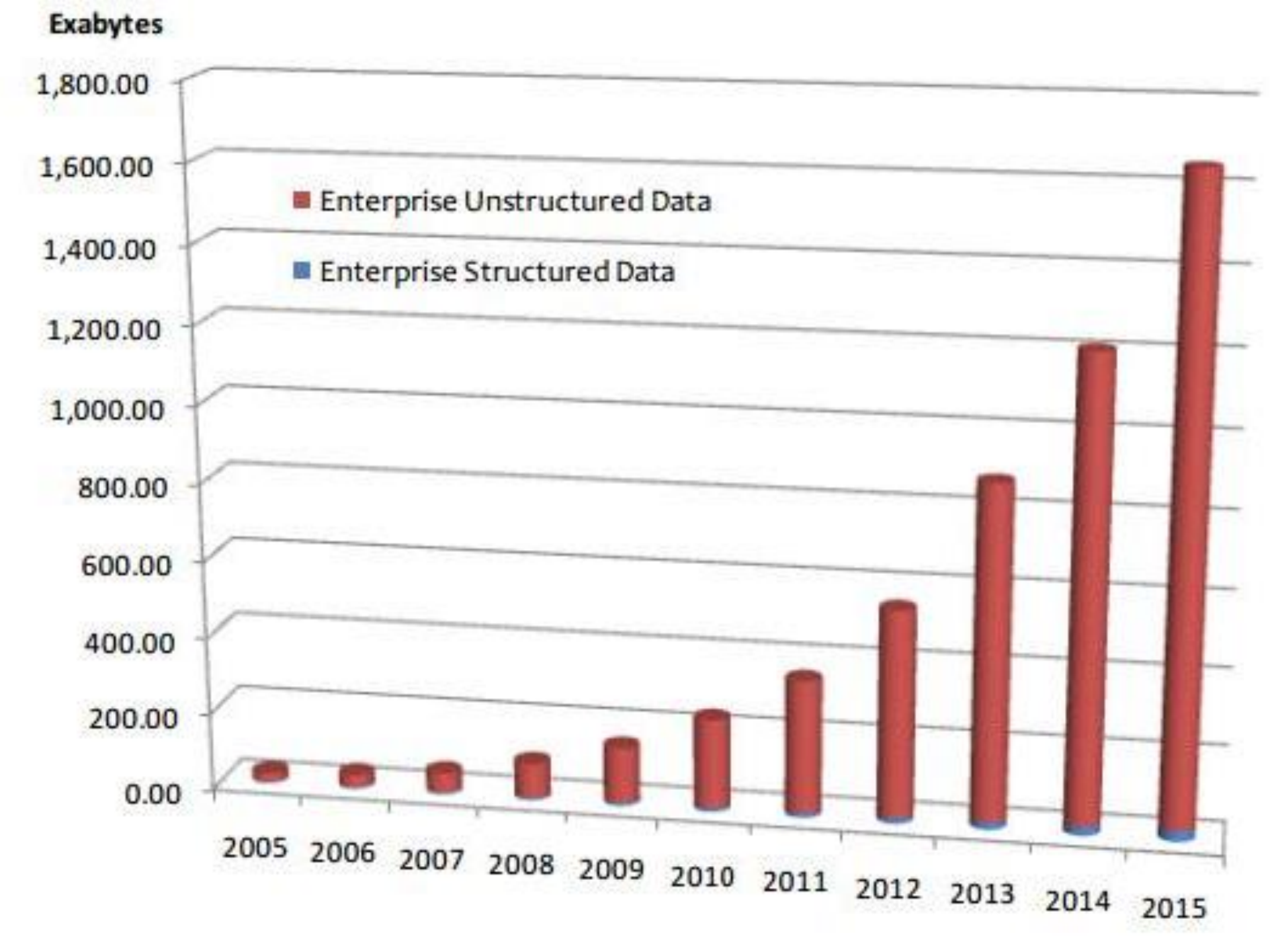
BIG DATA – ÞRÓUN, TÆKIFÆRI OG CRESS

Björn Þór Jónsson

Tölvunarfræðideild Háskólans í Reykjavík


IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
20,000+ NEW POSTS ON tumblr.
THE LARGEST SOCIAL READING PUBLISHING COMPANY
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
13,000+ iPhone APPLICATIONS DOWNLOADED
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
6,600+ NEW PICTURES ARE UPLOADED ON flickr
600+ NEW VIDEOS
50+ WORDPRESS DOWNLOADS
25+ HOURS TOTAL DURATION
70+ DOMAINS REGISTERED
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
60+ NEW BLOGS
168 MILLION EMAILS ARE SENT
694,445 SEARCH QUERIES
1,700+ Firefox DOWNLOADS
1,500+ BLOG POSTS
79,364 WALL POSTS
510,040 COMMENTS
Google
Google Search
GO-Globe.com


Exabytes
1,800.00
1,600.00
1,400.00
1,200.00
1,000.00
800.00
600.00
400.00
200.00
0.00
Enterprise Unstructured Data
Enterprise Structured Data
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

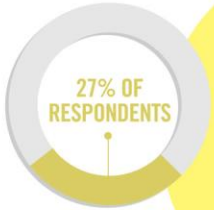
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions



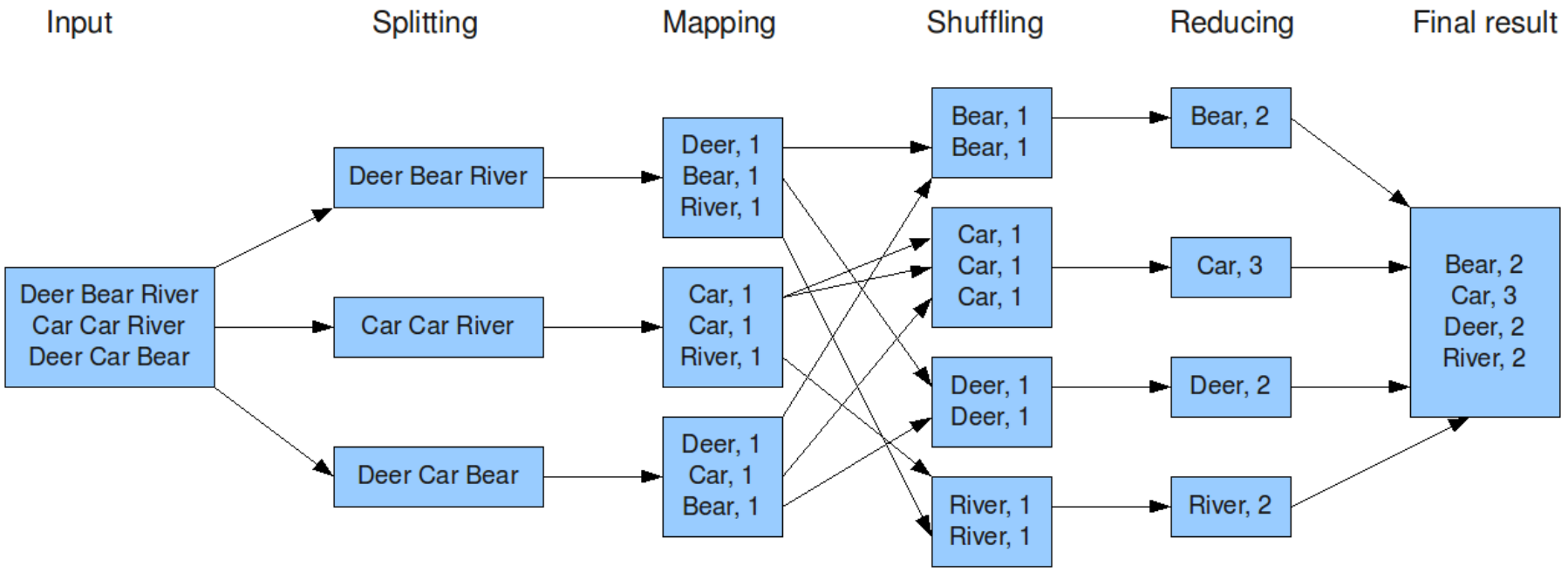
in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

The overall MapReduce word count process



The Big Data Landscape

Apps

Vertical

PREDICTIVE POLICING · Atigeo · MYRRIX · ellucian. · Placed.

Operational Intelligence

splunk> · loggly · sumologic

Ad/Media

METAMARKETS · collective[i] · rocketfuel · Recorded Future · bluefin · bloomreach · Quantivo · LuckySort · Media Science · TURN · DataXu

Data As A Service

factual. · FICO · kaggle · knoema beta · GNIP · DATASIFT · Windows Azure Marketplace · INRIX · LexisNexis · SPACE CURVE · LOCATE Everything Location

Business Intelligence

ORACLE | Hyperion · SAP · Business Objects · RJMetrics · Microsoft | Business Intelligence · IBM · COGNOS · birst · JASPERSOFT · pentaho · MicroStrategy · Autonomy · bime · DOMO · Chart.io · GoodData · ATTIVIO

Analytics and Visualization

+ableau · Palantir · OPERA · metaLayer · dataspora · TERADATA ASTER · centrifuge · SAS · TIBCO · KARMASPHERE · panopticon Real-Time Visual Data Analysis · Datameer · CIRRO · platfora · ClearStory · saffron · alteryx · visual.ly · AYATA · QlikView

Infrastructure

Analytics

Hortonworks · VERTICA An HP Company · MAPR · cloudera · INFOBRIGHT · ParAccel · EMC² · GREENPLUM. · NETEZZA · kognitio · DATASTAX · EXASOL · calpont

Operational

COUCHBASE · 10gen · HADAPT · TERRACOTTA · VoltDB · MarkLogic · INFORMATICA

As A Service

Qubole · amazon web services · Windows Azure · MORTAR · infochimps · Google BigQuery

Structured DB

ORACLE · MySQL · Microsoft SQL Server · PostgreSQL · IBM DB2 · SYBASE · memsql · TERADATA

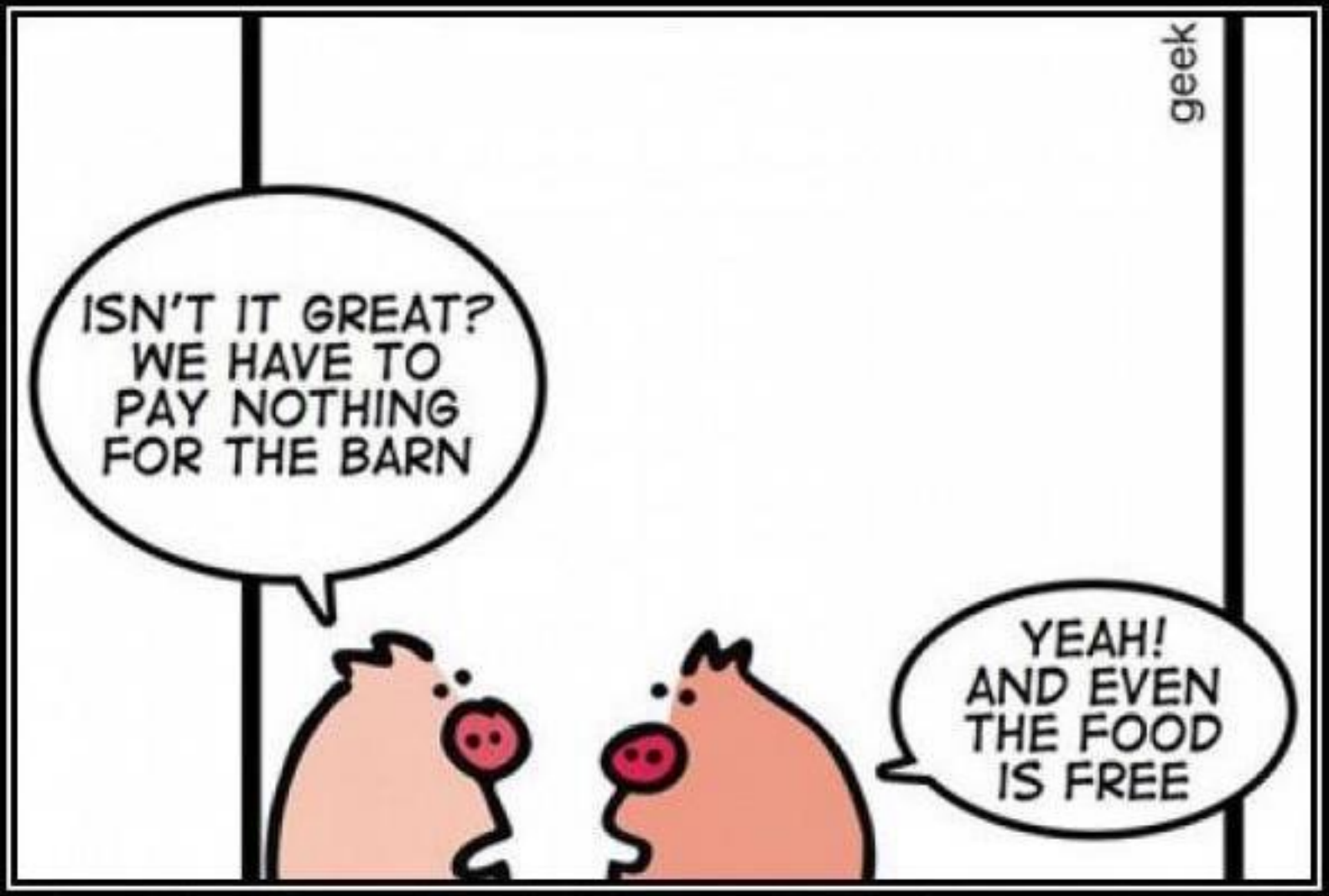
Technologies

hadoop · hadoop MapReduce · mahout · APACHE HBASE · Cassandra

dave@vcdave.com

bigdatalandcsape.com

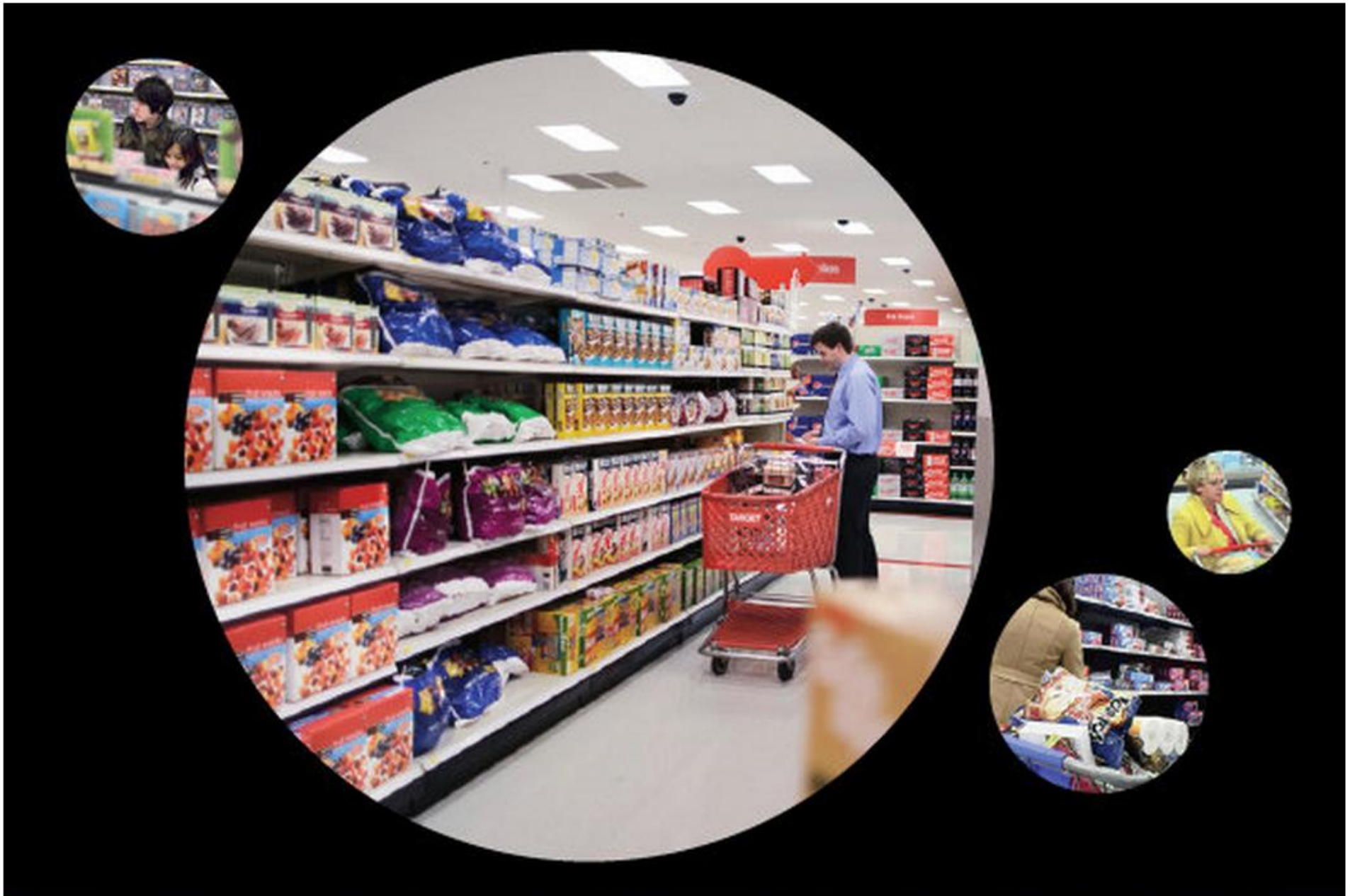

IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
20,000+ NEW POSTS ON tumblr.
THE LARGEST SOCIAL READING PUBLISHING COMPANY
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
13,000+ iPhone APPLICATIONS DOWNLOADED
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
6,600+ NEW PICTURES ARE UPLOADED ON flickr
600+ NEW VIDEOS
50+ WORDPRESS DOWNLOADS
25+ HOURS TOTAL DURATION
70+ DOMAINS REGISTERED
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
60+ NEW BLOGS
168 MILLION EMAILS ARE SENT
694,445 SEARCH QUERIES
1,700+ Firefox DOWNLOADS
1,500+ BLOG POSTS
79,364 WALL POSTS
510,040 COMMENTS
Google
Google Search
GO-Globe.com



FACEBOOK AND YOU

If you're not paying for it, you're not the customer. You're the product being sold.

How Companies Learn Your Secrets



TECH 2/16/2012 @ 11:02AM | 2,375,724 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

334 comments, 173 called-out + Comment Now + Follow Comments

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole — before Target freaked out and cut off all communications — about the clues to a customer's impending bundle of joy. Target assigns every customer a Guest ID number, tied to their credit card, name, or email address that becomes a bucket that stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources. Using that, Pole looked at historical buying data for all the ladies who had signed up for Target baby registries in the past. From the NYT:

Target has got you in its aim

Big Data Analytics: Making Government Data Work

"Big data" comes with many promises, but the data alone is not a silver bullet. True, it holds the potential for extracting business or mission intelligence and improving decision-making, but without the application of expert domain knowledge to give data contextual meaning, big data is nothing but a whole lot of dark figures.


CSC

A Briefing from GBC: Industry Insights
November 2012

Making Big Data Work for Government

"Big data" comes with many promises. Not the least of these is access to previously unaccounted information, "dark data," and large volumes of unstructured information. Many estimates place the amount of unstructured data worldwide at 90 percent of all data; and as the average federal agency stores 1.4 petabytes of data,[1] this represents a potentially vast store of previously untapped information for analysts.[2] But big data alone is not a silver bullet. True, it holds the potential for extracting business or mission intelligence and improving decision-making, but without the application of expert domain knowledge to give data contextual meaning, big data will be nothing but a whole lot of dark figures.

Currently, federal agencies cannot make use of all their data because they do not (or cannot afford to) employ enough data scientists – that is, experts who possess domain knowledge and can use big data technologies to ask the right questions and extract business or mission intelligence from vast pools of data. Making use of big data under these circumstances presents a unique challenge.

"Gartner estimates that big data initiatives in 2013 will total $34 billion,[3] and big data will drive $232 billion in spending over the next five years.[4]"

Subduing Big Data

The technological challenges of capturing, securing and managing the worldwide explosion of data are not insignificant (the amount of worldwide data currently measures about 2.7 zettabytes and is projected to double every two years[5]. Yet, corralling big data is as much an organizational challenge as a technological one. Industry development of big data technologies has largely bifurcated into two main functions: data discovery and the more traditional business intelligence.[6] Companies have been investing much time and money in the former.[7] Gartner estimates that big data initiatives in 2013 will total $34 billion,[8] and big data will drive $232 billion in spending over the next five years.[9]

Many organizations are trying to address the technology challenge of analyzing big data. The TechAmerica Foundation's Big Data Commission, for example, recently released a report that gives organizations guidelines for an effective big data program. Examination of these steps shows a decided focus not just on technical capabilities, but on organizational policies as well. The Commission recommends that CIOs take a "holistic approach to help guide the agency from an information management perspective and follow these key steps:

- Identify data and content that are critical to its mission
- Identify how, when, where and to whom information should be made available
- Determine appropriate data management, governance and security practices
- Identify and prioritize the information projects that deliver the most value.[10]

CSC

Government Business Council
Industry Insights

Government Big Data

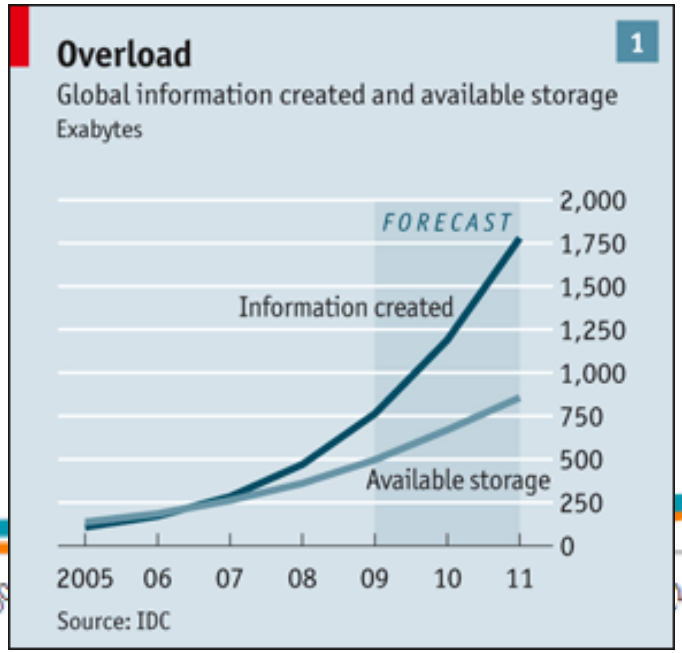
Currently, federal agencies cannot make use of all their government data because they do not (or cannot afford to) employ enough data scientists—that is, experts who possess domain knowledge and can use government big data analytic technologies to ask the right questions and extract business or mission intelligence from vast pools of data. Making use of big data under these circumstances presents a unique challenge.

Download the Big Data brief (PDF, 267KB)

Contact Us

Growth of Data vs. Growth of Data Analysts

Stored Data accumulating at 28% annual growth rate
Data Analysts in workforce growing at 5.7% growth rate



Data Scientist: The Hottest Job You Haven't Heard Of

By OnlineDegrees.com

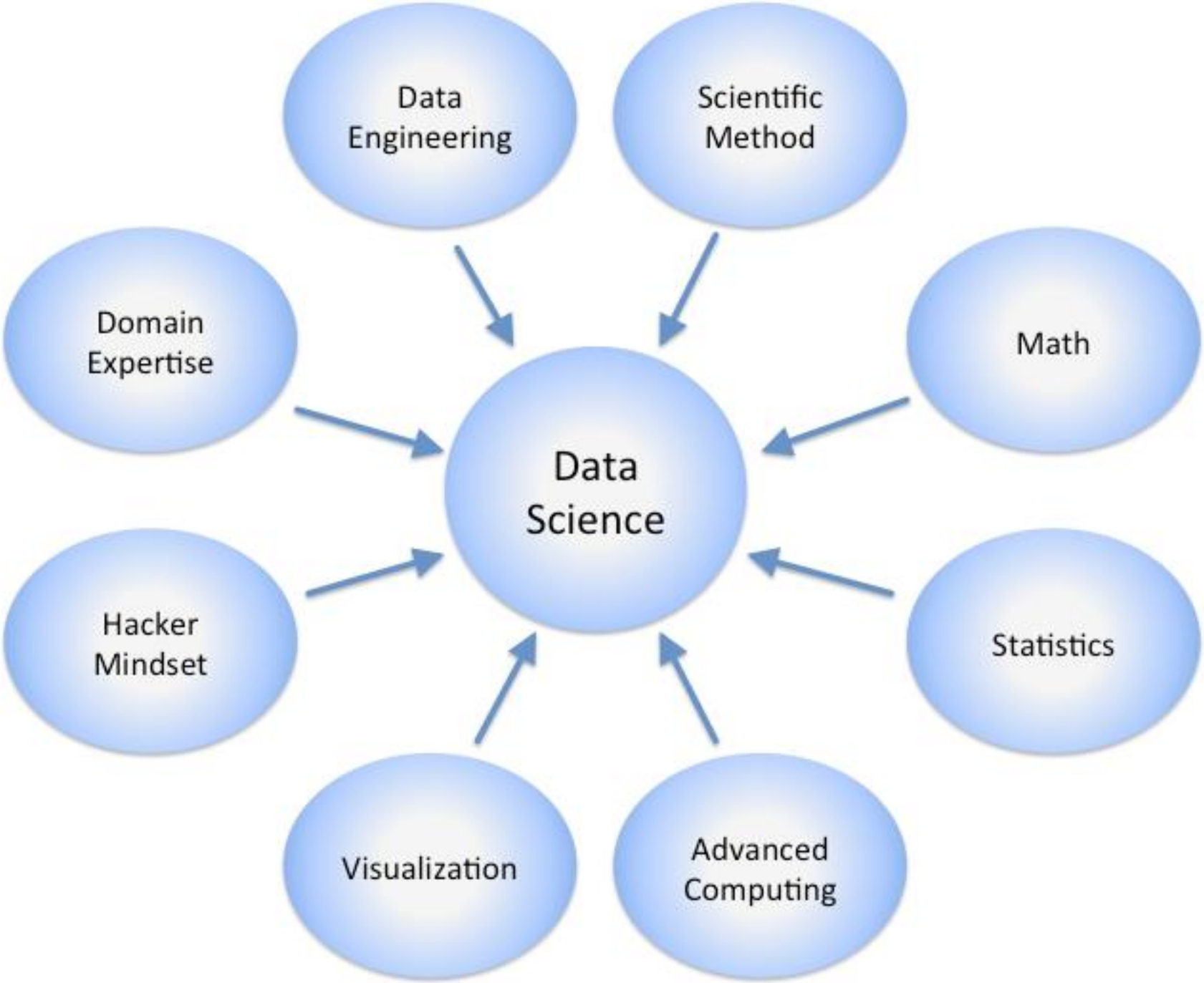
Posted Aug 10th 2011 @ 10:02AM

This article was originally featured on OnlineDegrees.com

By Maryalene LaPonsie

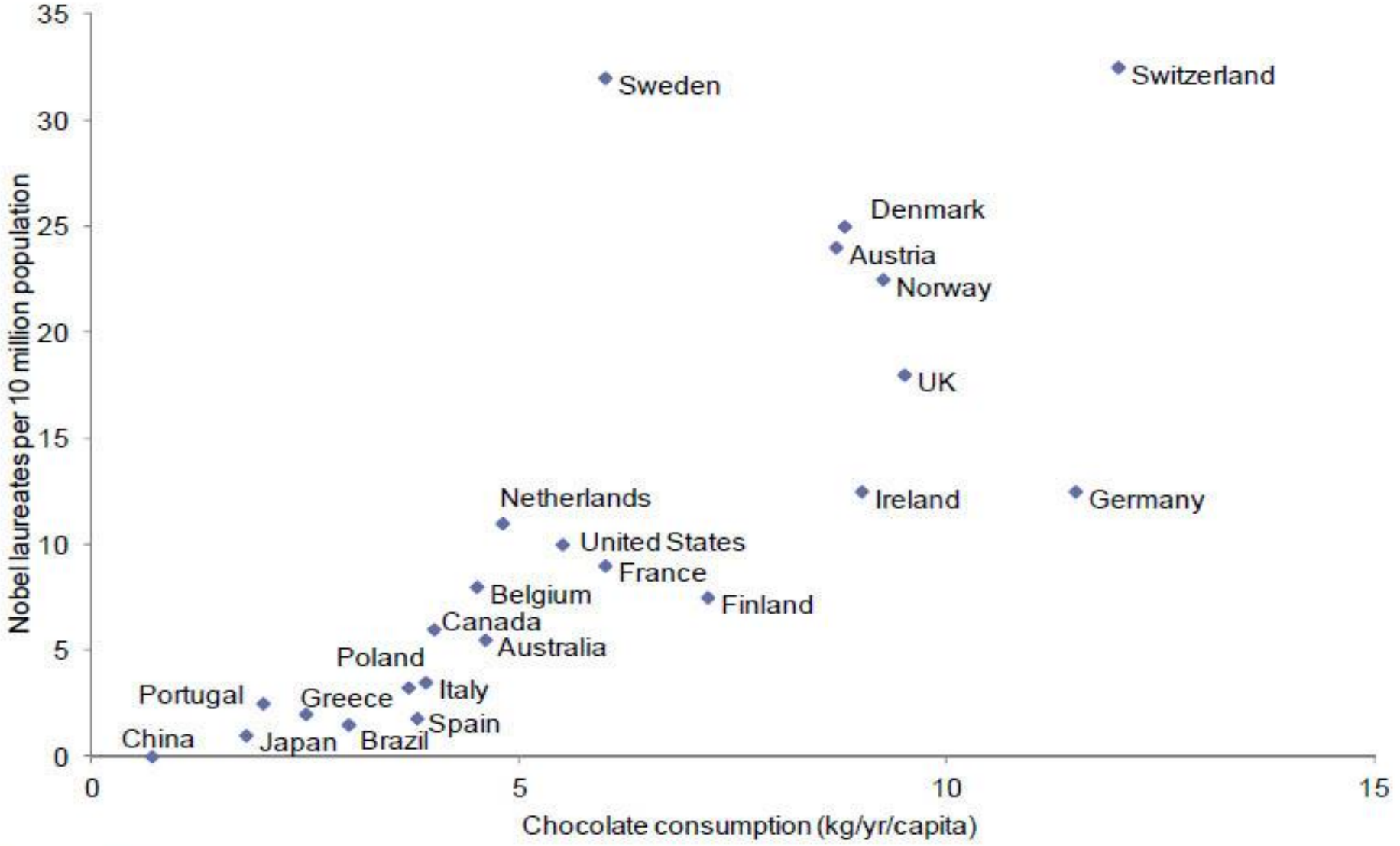
What has information overload done to us?

The search engine Bing would have us believe that we are all just a moment away from starting a food fight in the supermarket produce section.

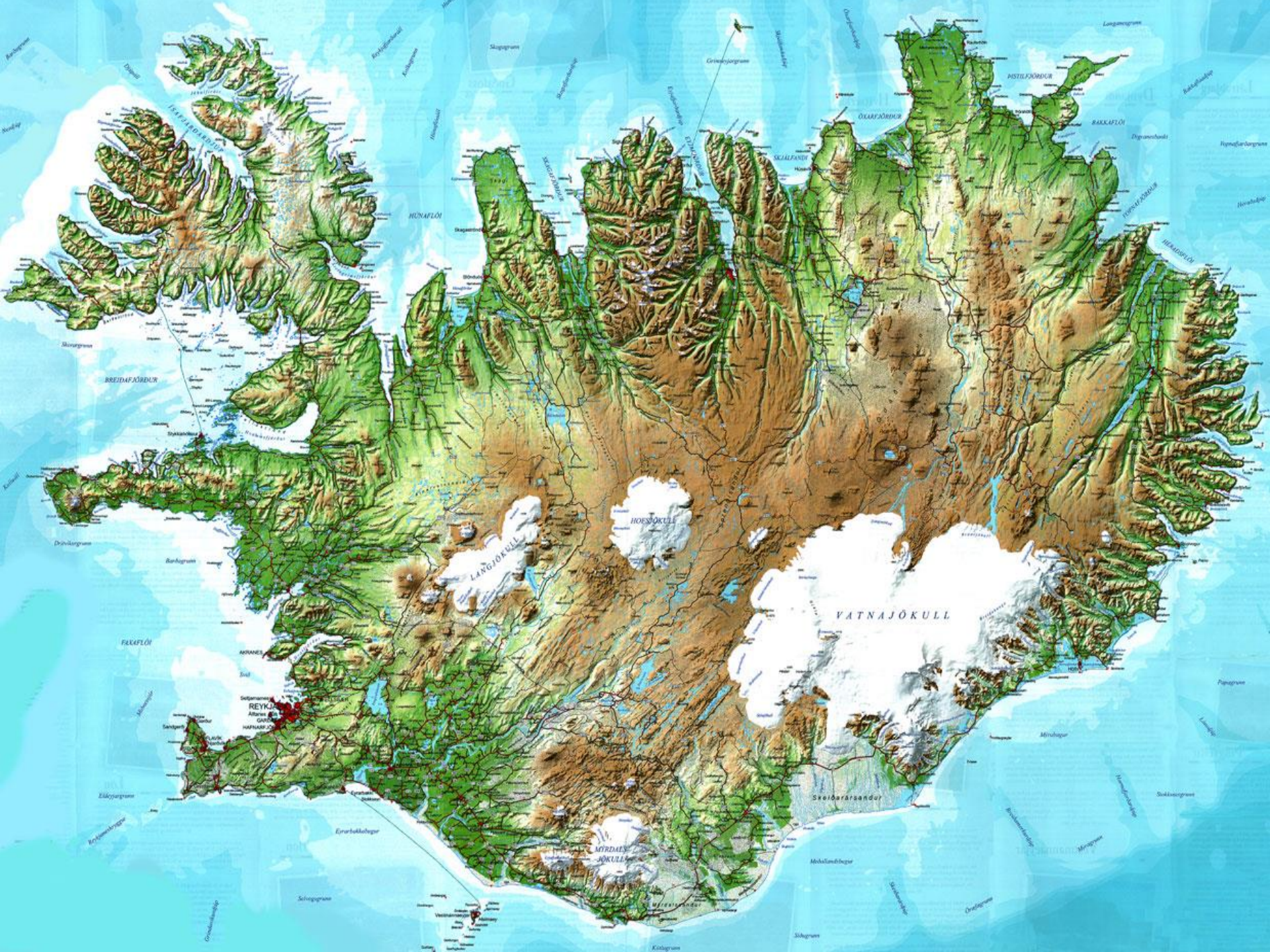

Data Engineering
Scientific Method
Domain Expertise
Math
Data Science
Hacker Mindset
Statistics
Visualization
Advanced Computing

Food for thought

Data as of 2012 or latest



Source: The New England Journal of Medicine


ÍSAFJARÐARDJÚP
HÚNAFLÓI
SKAGAFJÖRÐUR
EYJAFJÖRÐUR
SKJÁLFANDI
ÖXARFJÖRÐUR
ÞISTILFJÖRÐUR
RAKKAFLÓI
HERAÐSFLÓI
BREIÐAFJÖRÐUR
FAXAFLÓI
LANGJÖKULL
HOFSJÖKULL
VATNAJÖKULL
MÝRDALS-JÖKULL
Skeiðarársandur
AKRANES
REYKJAVÍK
HAFNARFJÖRÐUR
Eyrarbakkabugur
Meðallandsbugur
Mýrdalssandur

164	Guyana	784,894	July 1, 2010	0.011%	Annual official esti
165	Bhutan	744,910	March 9, 2014	0.0104%	Official population
166	Comoros	743,798	July 1, 2013	0.0104%	Official estimate
167	Montenegro	620,029	April 1, 2011	0.0087%	Final 2011 census
168	*Macau* (China)	598,200	September 30, 2013	0.0084%	Official estimate
169	Solomon Islands	581,344	July 1, 2013	0.0081%	Annual official esti
170	Western Sahara[18]	567,000	July 1, 2013	0.0079%	UN estimate
171	Luxembourg	537,000	December 31, 2012	0.0075%	Annual official esti
172	Suriname	534,189	August 13, 2012	0.0075%	Preliminary 2012 c
173	Cape Verde	491,875	June 16, 2010	0.0069%	Final 2010 census
174	Malta	416,055	November 20, 2011	0.0058%	Preliminary 2011 c
175	*Guadeloupe* (France)	405,739	January 1, 2013	0.0056%	Official annual esti
176	*Martinique* (France)	392,291	January 1, 2011	0.0055%	Official annual esti
177	Brunei	393,162	June 20, 2011	0.0055%	Preliminary 2011 c
178	Bahamas	351,461	May 3, 2010	0.0049%	Final 2010 census
179	Belize	349,728	July 1, 2013	0.0044%	Official estimate
180	Iceland	325,671	December 31, 2013	0.0045%	Quarterly official es
181	Maldives	317,280	July 1, 2010	0.0044%	Official estimate
182	Barbados	285,000	July 1, 2013	0.004%	Official estimate
183	*French Polynesia* (France)	268,270	August 22, 2012	0.0038%	Preliminary 2012 c
184	Vanuatu	264,652	July 1, 2013	0.0037%	Annual official esti
185	*New Caledonia* (France)	258,958	July 1, 2013	0.0036%	Annual official esti
186	*French Guiana* (France)	237,549	January 1, 2011	0.0032%	Official annual esti
187	*Mayotte* (France)	212,645	August 21, 2012	0.003%	2012 census result

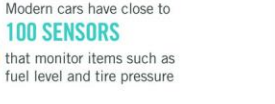
The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States



Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE
have cell phones

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

in one survey were unsure of how much of their data was inaccurate

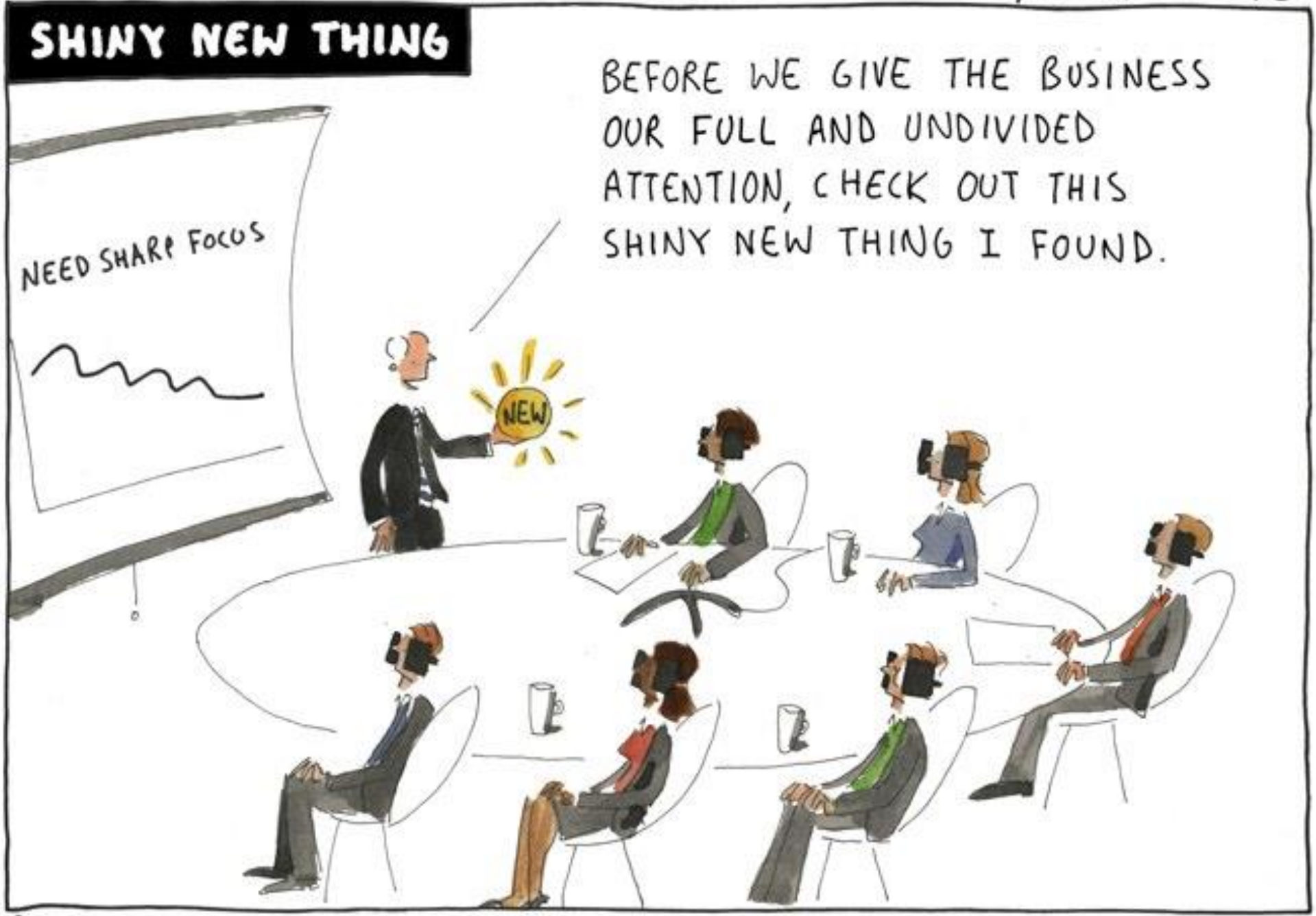
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

by Tom Fishburne

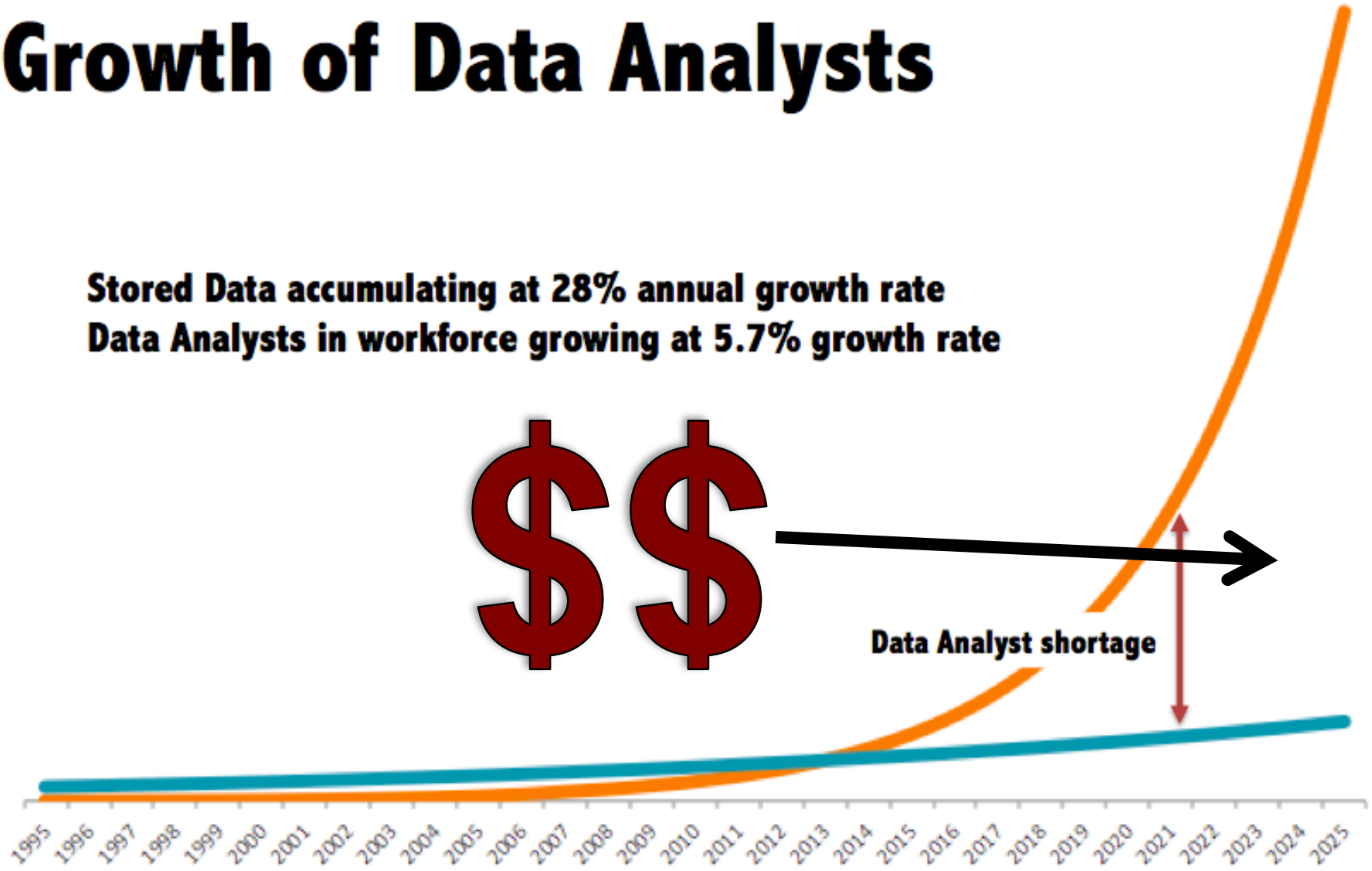
SHINY NEW THING




The Big Data Landscape
Apps
Vertical
Operational Intelligence
Ad/Media
Business Intelligence
Analytics and Visualization
Data As A Service
Infrastructure
Analytics
Operational
As A Serv
Technologies
Google Bi
dave@vcdave.com

Growth of Data vs. Growth of Data Analysts

Stored Data accumulating at 28% annual growth rate
Data Analysts in workforce growing at 5.7% growth rate



DATALA

HCI LAB

CRESS

Center for Research in Engineering Software Systems

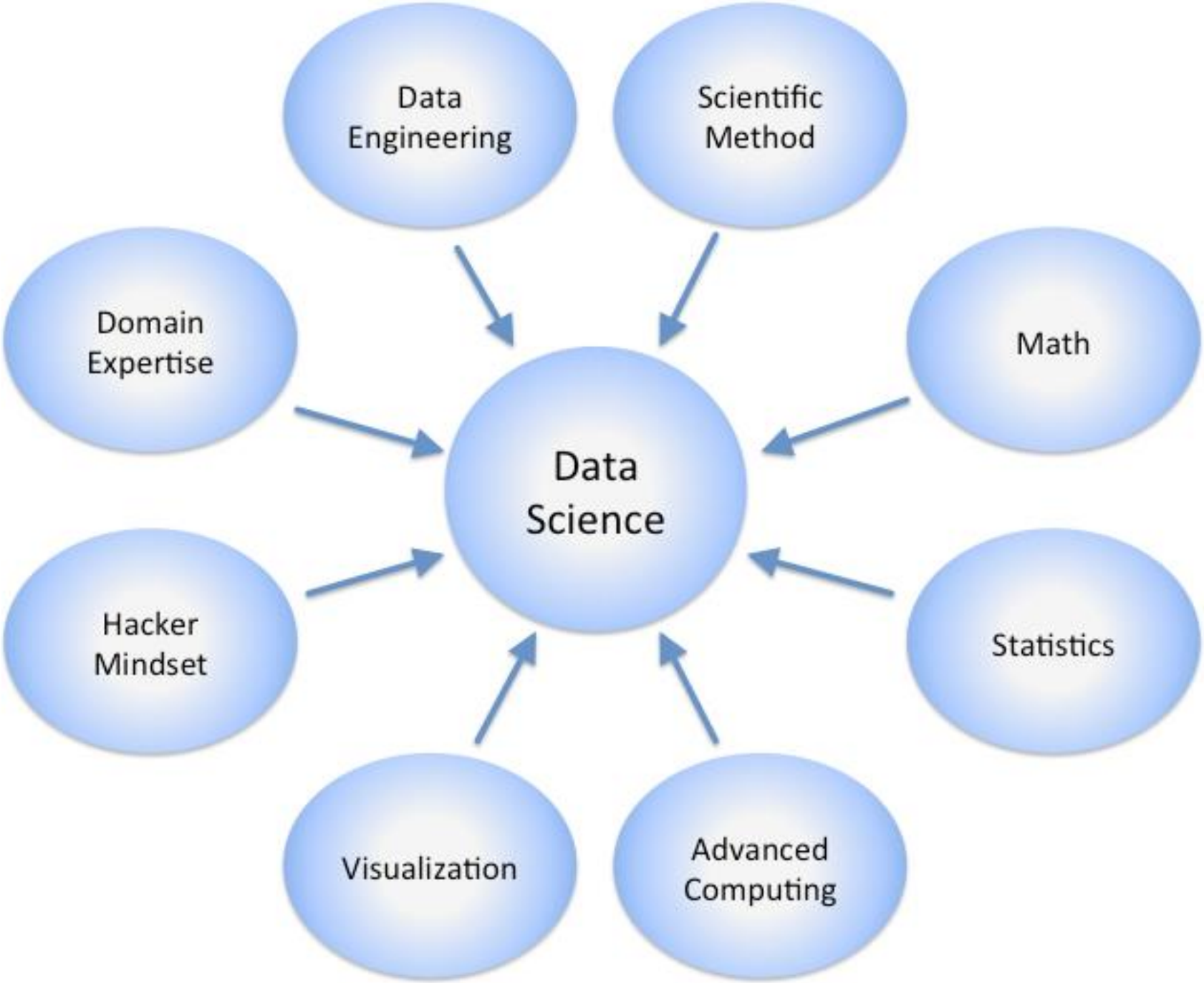
SYSLAB

ROSE

OPENING EVENT
THE CRESS RESEARCH CENTER AT REYKJAVIK UNIVERSITY
AUGUST 28, 2013

Dr. Michael J. Franklin
Siebel Professor of Computer Science, UC Berkeley
Director, Algorithms, Machines and People Laboratory (AMPLab)
Former founder and CTO of Truviso Inc.

Building Effective Industry/Academia Collaborations
The Berkeley Lab Model


Data Engineering
Scientific Method
Domain Expertise
Math
Data Science
Hacker Mindset
Statistics
Visualization
Advanced Computing

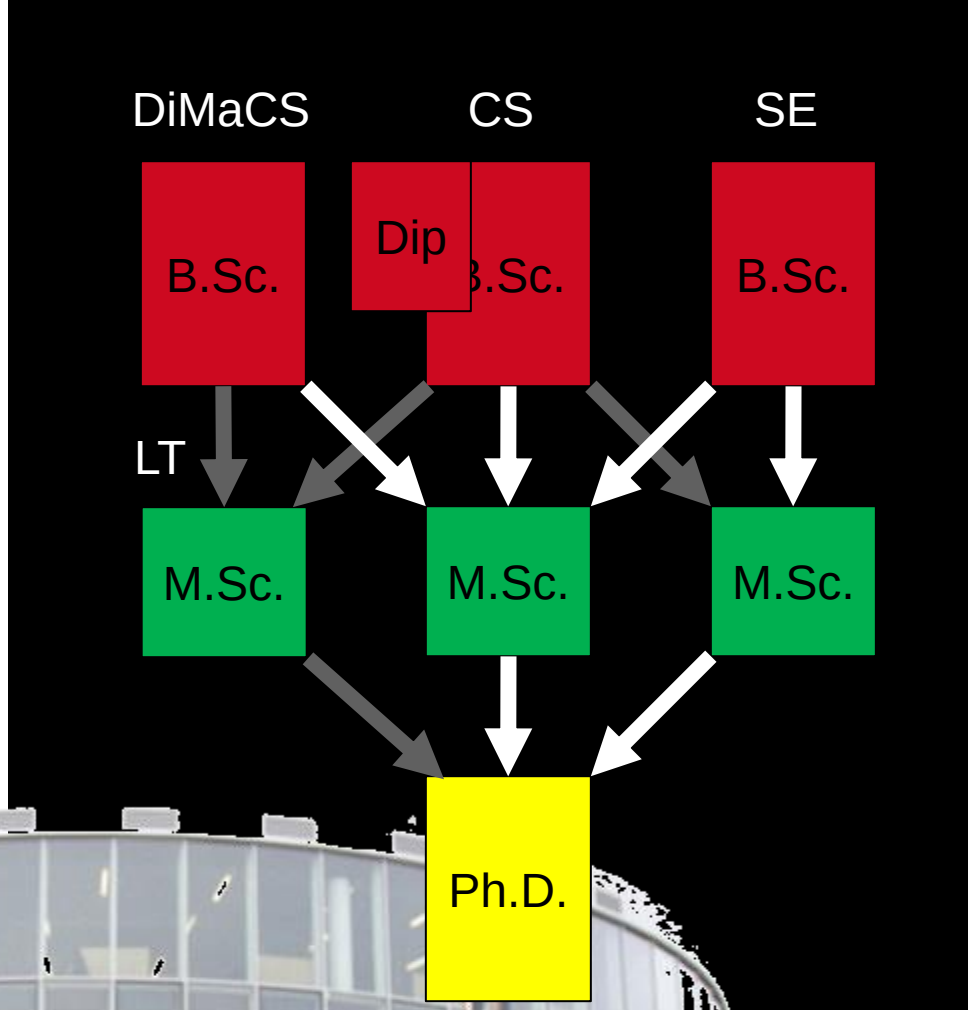
Vilja blása ungu fólki von í brjóst með áburðaverksmiðju

Aðalsteinn Kjartansson
adalsteinn@dv.is

16:56 › 27. febrúar 2014

Átta þingmenn Framsóknarflokksins hafa lagt fram tillögu til þingsályktunar um að ríkisstjórnin kanni hagkvæmni og möguleika þess að reisa áburðaverksmiðju í Helguvík eða Þorlákshöfn. Þetta á að vekja ungum Íslendingum von í brjósti um að stjórnvöld ætli sér að skapa þeim tækifæri og atvinnuöryggi í framtíðinni, að því er segir í greinargerð ályktunarinnar.

- BSc courses
  - Databases
  - Performance of Database Systems
  - Machine Learning
  - ...
- MSc courses:
  - Web Mining
  - *Analytics*
  - *Big Data*
  - ...



Sources

- http://jobs.aol.com/articles/2011/08/10/data-scientist-the-hottest-job-you-havent-heard-of/
- http://en.wikipedia.org/wiki/Data_science, http://en.wikipedia.org/wiki/MapReduce, http://en.wikipedia.org/wiki/Big_data, http://en.wikipedia.org/wiki/List_of_countries_by_population
- http://www.delphianalytics.net/wp-content/uploads/2013/04/GrowthOfDataVsDataAnalysts.png
- http://media.economist.com/images/20100227/201009SRC696.gif
- http://www.datasciencecentral.com/profiles/blogs/structured-vs-unstructured-data-the-rise-of-data-anarchy
- http://www.zerohedge.com/sites/default/files/images/user5/imageroot/2012/10-2/Food%20For%20Thoughts.jpg
- http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/
- http://www.theguardian.com/news/datablog/2012/mar/09/big-data-theory
- http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=3&hp&
- http://blogs-images.forbes.com/davefeinleib/files/2012/07/Big-Data-Trends.0031.png
- http://www.slideshare.net/4Neha/big-data-15681560
- http://www.mimul.com/pebble/default/images/blog/cloud/nosql_cap04.png
- http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg
- …